概率图模型笔记

宋佳欢

2019 年 10 月 23 日

目录

1 绪论

概率图的三个方面：
1. 表示：有向图（贝叶斯网络），无向图（马尔科夫网络），高斯图
2. 推断：精确推断，近似推断：确定性近似（变分推断），随机近似：MCMC
3. 学习: 参数学习，图结构学习

1.1 基础概念

高维随机变量的概率：

$$P(x_1, x_2, ..., x_p)$$

边缘概率：

$$P(x_i)$$

条件概率：

$$P(x_j|x_i)$$

加法法则（计算边缘概率）：

$$P(x_i) = \int P(x_1, x_2)dx_2$$

乘法法则（计算联合概率）：

$$P(x_1, x_2) = P(x_1) \cdot P(x_2|x_1)$$

链式法则（乘法的推广）：

$$P(x_1, x_2, \cdots, x_p) = \prod_{i=1}^{p} P(x_i|x_1, x_2, \cdots, x_{i-1}, x_{i+1}, \cdots, x_p)$$

贝叶斯定理：

$$P(x_2|x_1) = \frac{P(x_1, x_2)}{P(x_1)} = \frac{P(x_1, x_2)}{\int P(x_1, x_2)dx_2} = \frac{P(x_2)P(x_1|x_2)}{\int P(x_2)P(x_1|x_2)dx_2}$$

1.2 条件独立性

困境： $P(x_1, x_2, ..., x_p)$ 计算复杂，所以要简化：
1. 假设各个变量之间相互独立 (朴素贝叶斯)：$P(x|y) = \prod_{i=1}^{p} P(x_i|y)$
2. 现实性每个变量之间多少是有关联的，所以条件再放松一点，那就是马尔可夫性，

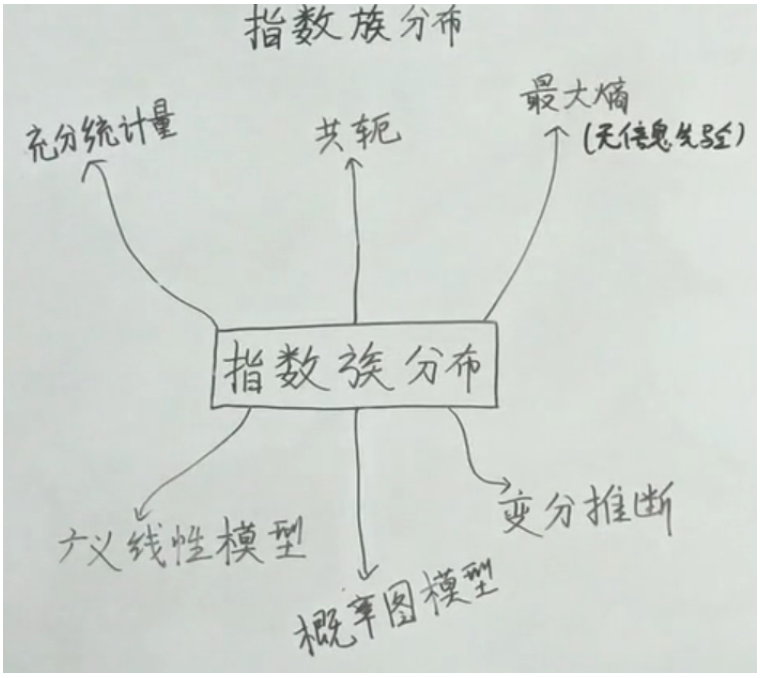
$$x_j \perp x_{i+1}|x_i, \quad j < i$$

3. 再推广，**条件独立性**假设:f

$$x_A \perp x_B|x_C, \quad x_A, x_B, x_C\text{是集合，且不相交}$$

条件独立性使用图来表示，在图上赋予概率的意义，使得图能表达条件独立性。

2 指数族分布



指数族分布的形式:

$$P(x|\eta) = h(x)exp(\eta^T\varphi(x) - A(\eta))$$

其中 η 为参数向量。$\varphi(x)$ 为充分统计量（给了充分统计量就能描述分布了）。$A(\eta)$：log partition function(对数配分函数)。相当与归一化因子，如：

$$P(x|\theta) = \frac{1}{Z}\hat{P}(x|\theta)$$

两边对 x 积分，化简得到：$Z = \int \hat{P}(x|\theta)dx$

2.1 高斯分布指数族形式

$$P(X|\theta) = \frac{1}{\sqrt{2\pi}\sigma}exp\big(-\frac{(x-\mu)^2}{2\sigma^2}\big)$$

将 x 与参数分开：

$$\begin{aligned} P(X|\theta) &= \frac{1}{\sqrt{2\pi\sigma^2}}exp\big(-\frac{1}{2\sigma^2}(x^2 - 2\mu x + \mu^2)\big) \\ &= exp[log(2\pi\sigma^2)^{-\frac{1}{2}}]exp\{-\frac{1}{2\sigma^2}(x^2 - 2\mu x) - \frac{\mu^2}{2\sigma^2}\} \\ &= exp\{\underbrace{(\frac{\mu}{\sigma^2} - \frac{1}{2\sigma^2})}_{\eta^T}\underbrace{\begin{pmatrix} x \\ x^2 \end{pmatrix}}_{\varphi(x)}\underbrace{-(\frac{\mu}{2\sigma^2} + \frac{1}{2}log2\pi\sigma^2)}_{A(\eta)}\} \end{aligned}$$

$$\eta = \begin{pmatrix} \eta_1 \\ \eta_2 \end{pmatrix} = \begin{pmatrix} \frac{\mu}{\sigma^2} \\ -\frac{1}{2\sigma^2} \end{pmatrix}$$

$$A(\eta) = -\frac{\eta_1^2}{4\eta_2} + \frac{1}{2}log(-\frac{\pi}{\eta_2})$$

2.2 对数分配函数与充分统计量之间的关系

即 $\varphi(x)$ 与 $A(\eta)$ 之间的关系。

$$\begin{aligned} P(x|\eta) &= h(x)exp(\eta^T\varphi(x) - A(\eta)) \\ &= \frac{1}{exp(A(\eta))}h(x)exp(\eta^T\varphi(x)) \end{aligned}$$

利用概率密度函数积分为 1 的性质，两边积分化简后得到：

$$exp(A(\eta)) = \int h(x)exp(\eta^T\varphi(x))dx$$

关于 η 求导：

$$exp(A(\eta))A'(\eta) = \frac{\partial}{\partial\eta}\Big(\int h(x)exp(\eta^T\varphi(x))dx\Big)$$

$$exp(A(\eta))A'(\eta) = \int h(x)exp(\eta^T\varphi(x)) \cdot \varphi(x)dx$$

$$A'(\eta) = \frac{\int h(x)exp(\eta^T\varphi(x)) \cdot \varphi(x)dx}{exp(A(\eta))}$$

$$A'(\eta) = \int \underbrace{h(x)exp\Big(\eta^T\varphi(x) - A(\eta)\Big)}_{P(x|\eta)} \cdot \varphi(x)dx$$

$$A'(\eta) = \int P(x|\eta) \cdot \varphi(x)dx = E_{P(x|\eta)}\Big[\varphi(x)\Big]$$

可证 $A(\eta)$ 的二阶导可得为 $\varphi(x)$ 的方差（这里没有证），所以有：

$$A'(\eta) = E_{P(x|\eta)}\Big[\varphi(x)\Big]$$

$$A''(\eta) = Var\Big[\varphi(x)\Big]$$

因为方差为正数，所以 $A(\eta)$ 是凸函数。

2.3 极大似然估计与充分统计量

上述讨论并没有引入数据，是根据公式推导的性质。下面对参数 η 进行极大似然估计，假设有样本集 $D = x_1, x_2, \cdots, x_N$，则：

$$\begin{aligned} \eta_{MLE} &= argmax logP(D|\eta) \\ &= argmax log\prod_{i=1}^{N} P(x_i|\eta) \\ &= argmax \sum_{i=1}^{N} logP(x_i|\eta) \\ &= argmax \sum_{i=1}^{N} log\Big[h(x)exp(\eta^T\varphi(x) - A(\eta))\Big] \\ &= argmax \sum_{i=1}^{N} \Big(\eta^T\varphi(x) - A(\eta)\Big) \quad \text{(去除无关的变量)} \end{aligned}$$

将上式关于 η 求导并置为 0：

$$\begin{aligned}\frac{\partial}{\partial\eta}\sum_{i=1}^{N}\Big(\eta^T\varphi(x)-A(\eta)\Big)&=\sum_{i=1}^{N}\frac{\partial}{\partial\eta}\Big(\eta^T\varphi(x)-A(\eta)\Big)\\&=\sum_{i=1}^{N}\Big(\varphi(x_i)-A'(\eta)\Big)\\&=\sum_{i=1}^{N}\varphi(x_i)-NA'(\eta)\\&=0\end{aligned}$$

得到：

$$A'(\eta)=\frac{1}{N}\sum_{i=1}^{N}\varphi(x_i)$$

求参数 η, 再加一步，求 A 的导数的反函数：

$$\eta=A'^{(-1)}(\eta)$$

可以得到结论：可以仅保留充分统计量求出参数，样本就可以扔了。

2.4 最大熵角度

在满足已知事实条件下（约束），最大熵的分布就是我们要的分布。若没有任何已知的情况下，均匀分布的熵最大。

经验分布：对已知样本的描述。

有样本 $D=x_1,x_2,\cdots,x_N$, 则经验分布：

$$\hat{P}(X=x)=\frac{count(x)}{N}$$
$$E_{\hat{p}}[x],\quad Var_{\hat{p}}[x]$$

设 $f(x)$ 是任意关于 x 的函数，已知事实就描述为：

$$E_{\hat{p}}[f(x)]=\Delta$$

最大熵原理：

$$min\quad\sum_x p(x)logp(x)$$
$$s.t.\sum_x p(x)=1,\quad E_p[f(x)]=E_{\hat{p}}[f(x)]=\Delta$$

我们将 $f(x)$ 推广到多维情形：

$$f(x)=\begin{pmatrix}f_1\\f_2\\\vdots\\f_Q\end{pmatrix},\Delta=\begin{pmatrix}\Delta_1\\\Delta_2\\\vdots\\\Delta_Q\end{pmatrix}$$

利用拉格朗日乘子法求解最大熵模型：

$$L(p,\lambda_0,\lambda)=\sum_{i=1}^{N}p(x_i)logp(x_i)+\lambda_0\Big(1-\sum_{i=1}^{N}p(x_i)\Big)+\lambda\Big(\Delta-E_p[f(x)]\Big)$$

求偏导置为 0：

$$\frac{\partial L}{\partial p(x_i)}=logp(x_i)+1-\lambda_0-\lambda^T f(x_i)=0$$

$$logp(x)=\lambda^T f(x)+\lambda_0-1$$

$$p(x)=exp\{\underbrace{\lambda^T f(x)}_{\eta^T\varphi(x)}-\underbrace{(\lambda_0+1)}_{A(\eta)}$$

结论：由最大熵原理推出来的模型就是指数族分布。

2.5 共轭分布

共轭分布的概念：

当某个分布的先验分布碰上了某种似然函数，其后验分布就会和先验分布的类型相同：

$$\underbrace{P(\beta|X)}_{\text{后验分布}}\propto\underbrace{P(X|\beta)}_{\text{似然函数}}\underbrace{P(\beta)}_{\text{先验分布}}$$

当先验与似然都为指数族分布（假设了似然函数为一维）：

$$P(\beta|X)\propto\underbrace{h(X)exp\{\beta\varphi(X)-A_l(\beta)\}}_{P(X|\beta)}\cdot\underbrace{h(\beta)exp\{\alpha^T\varphi(\beta)-A(\alpha)\}}_{P(\beta)}$$

其中 $h(X)$ 和 $A(\alpha)$ 对于 β 来说是常数, 下面可以忽略。我们下面做这样的假设：

$$\varphi(\beta)=\begin{pmatrix}\beta\\-A_l(\beta)\end{pmatrix}\quad\alpha=\begin{pmatrix}\alpha_1\\\alpha_2\end{pmatrix}$$

我们得到：

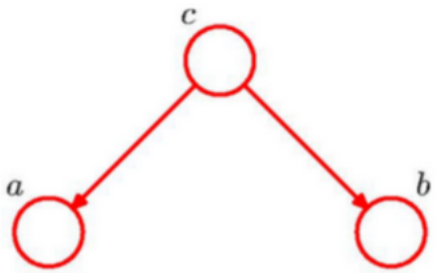
$$\begin{aligned}P(\beta|X)&\propto h(\beta)exp\{\beta\varphi(X)-A_l(\beta)+\alpha_1\beta-\alpha_2A_l(\beta)\}\\&=h(\beta)exp\{(\varphi(X)+\alpha_1)\beta-((1+\alpha_1)A_l(\beta))\}\\&=h(\beta)exp\{\underbrace{(\hat{\alpha_1},\hat{\alpha_2})}_{\text{新的参数}}\begin{pmatrix}\beta\\-A_l(\beta)\end{pmatrix}\}\end{aligned}$$

如果似然函数的对数配分函数与先验分布的充分统计量的第二部分相同，那么先验与后验属于同一分布（共轭）。所以指数族分布的似然函数必定能找到一个使得后验分布共轭的先验分布。

3 贝叶斯网络

3.1 贝叶斯网络的三种结构的独立性

1.tail to tail



计算三个变量的联合概率：

$$\text{因子分解:}\quad P(a,b,c)=P(c)P(a|c)P(b|c)$$

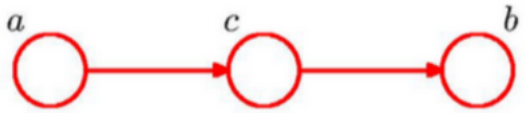
$$\text{链式法则:}\quad P(a,b,c)=P(c)P(a|c)P(b|a,c)$$

可得：

$$P(b|c)=P(b|a,c)\Longrightarrow a\perp b|c$$

该图的条件独立性：**若 c 被观测，则路径阻塞**（图论的说法，阻塞意味着独立）

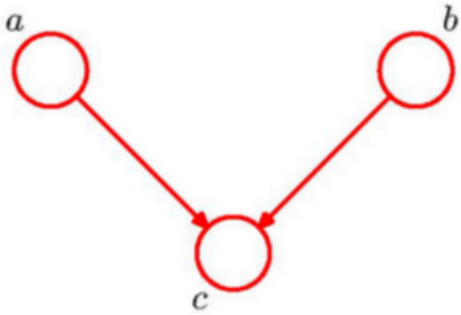
2.head to tail



f

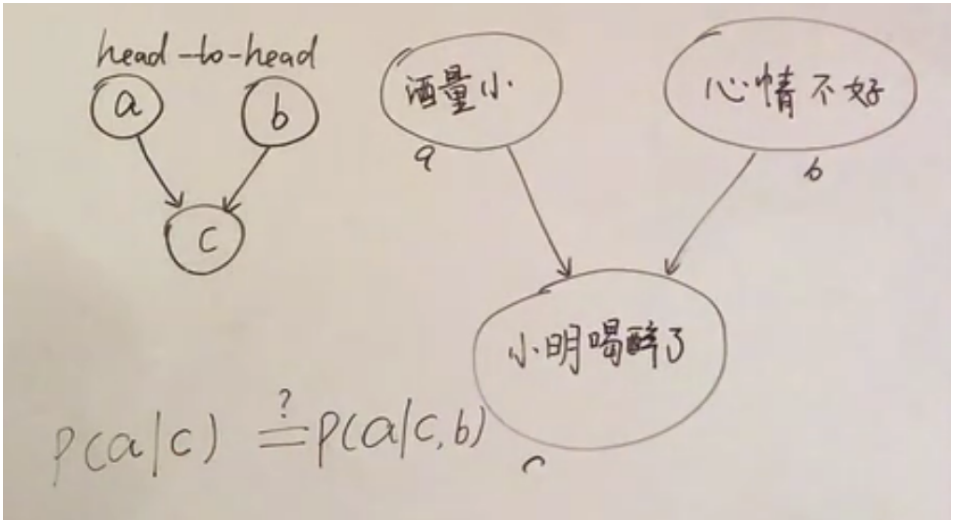
若 c 被观测，则路径阻塞,$a\perp b|c$

3.head to head



若 c 被观测，则路径是通的，即 a，b 之间不独立

一个例子：



得知小明喝醉，推小明酒量小的概率 $P(a|c)$ 应该是比较大的。

得知小明喝醉且心情不好，推小明酒量小的概率 $P(a|c,b)$ 就比上一种情况小了。因此 a，b 是相关的。

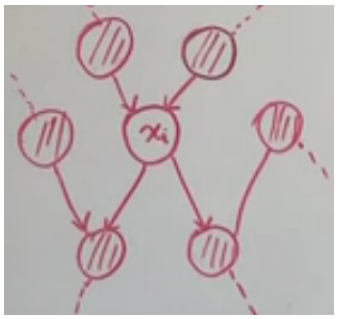
3.2 D 划分（d-Separation）

1. 集合 A,B,C，通过 tail to tail 和 head to tail 结构连接，A 到 B 的路径上的节点都必须在 C 的内部。

2. 集合 A,B,C，通过 head to head 结构连接，A 到 B 的路径上的节点以及其后继节点都不能在 C 的内部。

满足上述两个要求，则 $A \perp B|C$。即满足全局马尔可夫性。

以下图为例，计算边缘概率 $P(x_i|x_1, x_2, \cdots, x_{i-1}, x_{i+1}, \cdots, x_p)$:



$$
\begin{aligned}
P(x_i|x_1, x_2, \cdots, x_{i-1}, x_{i+1}, \cdots, x_p) &= \frac{P(X)}{P(x_1, x_2, \cdots, x_{i-1}, x_{i+1}, \cdots, x_p)} \\
&= \frac{P(X)}{\int P(X)dx_i} \\
&= \frac{\prod_{j=1}^{p} P(x_j|x_{parent(j)})}{\int \prod_{j=1}^{p} P(x_j|x_{parent(j)})dx_i}
\end{aligned}
$$

$x_{parent(j)}$表示x_i的父节点们

可见 x_i 的边缘概率只与和它相关的一些节点有关，即上图中的打阴影的节点,x_i 周围的一圈又叫马尔可夫毯。

4 Markov 网络

全局马尔可夫性：从节点集 A 中的节点到 B 中的节点的路径，都经过节点集 C 中的节点，则 $A \perp B|C$。

局部马尔可夫性：$a \perp \{全集 - a - 邻居\}|邻居$。

成对马尔可夫性：$x_i \perp x_j|\{全集 - x_i - x_j\}$

团: 图中节点的一个子集，其中任意两个节点有互相连接。

极大团：在极大团中再加入一个节点就不够成团。

因子分解 (极大团的势函数相乘)：

$$P(X) = \frac{1}{Z}\prod_{i=1}^{k}\psi(x_{C_i})$$

5 推断 inference

目的：

求联合概率：$P(x_1, x_2, ..., x_p)$

求边缘概率：$P(x_i) = \sum_{x_1}\cdots\sum_{x_{i-1}}\sum_{x_{i+1}}\cdots\sum_{x_p}P(X)$

求条件概率：$P(x_A|x_B)$

最大化后验概率：$\hat{Z} = arg\max_Z P(Z|X) \propto argmaxP(Z, X)$

精确推断：

变量消去 (variable elimanation)，信念传播 (belief propagation)(Sum-product 针对树结构)，junction tree(普通图结构)。

近似推断：

loop belief propagation(有环图)，import sampling, MCMC, Variational Inference。

5.1 变量消去 (variable elimanation)

以马尔可夫链为例:

$$ⓐ \to ⓑ \to ⓒ \to ⓓ$$

$$P(d) = \sum_{a,b,c}P(a, b, c, d) = \sum_{a,b,c}P(a)P(b|a)P(c|b)P(d|c)$$

用上面的公式直接计算边缘概率，计算量会随着可选状态数增加而指数增加，因此需要化简。可以看到，关于 a 求和时，求和符号内的有些项式与 a 无关的，可以提出去，可以写成：

$$\begin{aligned} P(d) &= \sum_{a,b,c} P(a)P(b|a)P(c|b)P(d|c) \\ &= \sum_{b,c} P(c|b)P(d|c) \underbrace{\sum_{a} P(a)P(b|a)}_{\varphi_a(b)} \\ &= \sum_{c} P(d|c) \underbrace{\sum_{b} P(c|b)\varphi_a(b)}_{\varphi_b(c)} \\ &= \varphi_c(d) \end{aligned}$$

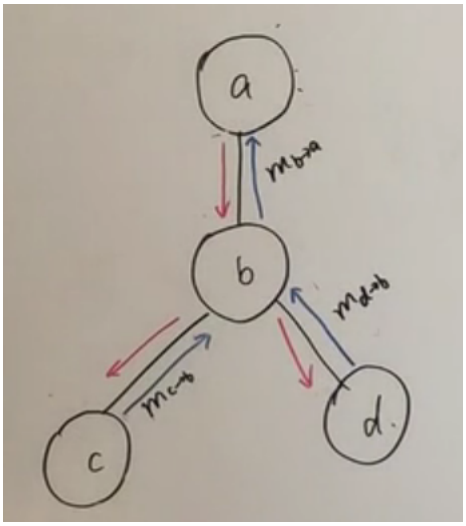
（乘法分配率）

5.2 信念传播 (belief propagation)

如下图所示，如果计算无向图中节点 a 的概率：

$$P(a) = \sum_{b,c,d} P(a,b,c,d)$$

$$P(a,b,c,d) = \frac{1}{Z}\varphi_a(a)\varphi_b(b)\varphi_c(c)\varphi_d(d)\varphi_{a,b}(a,b)\varphi_{b,c}(b,c)\varphi_{b,d}(b,d) \quad \text{（四个点，三条边）}$$



可以先计算与 c 或 d 相关的项：

$$m_{c\to b}(b) = \sum_{c} \varphi_c(c)\varphi_{b,c}(b,c)$$

$$m_{d\to b}(b) = \sum_{d} \varphi_d(d)\varphi_{b,d}(b,d)$$

再计算 $m_{b\to a}$：

$$m_{b\to a}(a) = \sum_{b} \varphi_b(b)\varphi_{a,b}(a,b)m_{c\to b}(b)m_{d\to b}(b)$$

最后得到 a 的边缘概率:

$$P(a) = \varphi_a(a) m_{b\to a}(a)$$

跳出这个例子，得到更加普遍的计算方式：

$$m_{j\to i}(x_i) = \sum_{x_j} \varphi_{ij}\, \varphi_j \underbrace{\prod_{k\subset NB(j)-i} M_{k\to j}(x_j)}_{belief(i)}$$

$$P(x_i) = \varphi_i \prod_{k\subset NB(i)} M_{k\to i}(x_i)$$

其中 $NB(j) - i$ 表示除了 x_i 以外，x_j 的所以邻居节点的集合。

还以上图为例，上面公式可以理解成每个节点都在收集关于 a 的信息，b 从 c,d 中收集关于 a 的所以信息，然后再送到 a。beilef(b) 即表示为在 b 到 a 这条路径上 b 所能提供的信息量。

BP 算法所做的事情就是遍历整个图，计算所有的 m_{ij}, 然后在到处边缘概率。分为下面几步：

1. 选个一个根节点
2. 收集信息
3. 发送信息

5.3 Max-product Algorithm

该算法可看成 BP 的改进，Viterb 的推广。

重新定义 $m_{j\to i}$：

$$m_{j\to i}(x_i) = \max_{x_j} \varphi_{ij}\varphi_j \prod_{k\subset NB(j)-i} M_{k\to j}(x_j)$$

找到一个最优的 $\max_{x_j}$ 的状态，使得 $m_{j\to i}$ 达到最大。

6 变分推断 (Variational Inference)

6.1 引入

贝叶斯推断（求后验分布）：

$$\underbrace{P(\theta|X)}_{\text{后验概率}} = \frac{P(X|\theta)P(\theta)}{\underbrace{P(X)}_{\int_\theta P(X|\theta)P(\theta)d\theta}}$$

贝叶斯决策（计算新样本的后验概率）：假设给定了 N 个观测数据 X，新的样本 $\tilde{x}$，求 $P(\tilde{x}|X)$:

引入参数 θ:

$$P(\tilde{x}|X) = \int_\theta P(\tilde{x}, \theta|X) d\theta$$

$$
\begin{aligned}
P(\tilde{x}|X) &= \int_{\theta} P(\tilde{x}|\theta, X)\underbrace{P(\theta|X)}_{\text{后验}}d\theta \\
&= \int_{\theta} P(\tilde{x}|\theta)P(\theta|X)d\theta \quad (\text{给定了}\theta, X\text{与}\tilde{x}\text{无关}) \\
&= E_{P(\theta|X)}[P(\tilde{x}|\theta, X)]
\end{aligned}
$$

所以可以用推断求得的后验去预测新的样本。

问题主要集中在求后验分布（推断），如果非常简单，就可以通过精确推断求解，但是常常参数或隐变量的空间维度特别高，很难取求解那个积分，所以可通过近似推断来求解。

6.2 推导

X: observed data

Z: latent variable + paramter

(X,Z): complete data

$$
\begin{aligned}
lnP(X) &= lnP(X, Z) - lnP(Z|X) \\
&= ln\left(\frac{P(X, Z)}{q(Z)}\right) - ln\left(\frac{P(Z|X)}{q(Z)}\right) \\
&= lnP(X, Z) - lnq(Z) - ln\left(\frac{P(Z|X)}{q(Z)}\right)
\end{aligned}
$$

两边求关于分布 $q(X)$ 的期望：

$$
lnP(X) = \underbrace{\int_Z lnP(X, Z)q(Z)dZ - \int_Z lnq(Z)q(Z)dZ}_{ELBO} \underbrace{- \int_Z ln\left(\frac{P(Z|X)}{q(Z)}\right)q(Z)dZ}_{KL\text{散度}(\geq 0)}
$$

所以 $lnP(x)$ 是 ELBO 的上界。

从另一角度来推导：

$$
\begin{aligned}
lnP(X) &= ln\Big(\int_Z P(X, Z)dZ\Big) \\
&= ln\Big(\int_Z \frac{P(X, Z)}{q(Z)} \cdot q(Z)dZ\Big) \\
&= lnE_{q(Z)}\Big[\frac{P(X, Z)}{q(Z)}\Big] \\
&\geq E_{q(Z)}\Big[ln\frac{P(X, Z)}{q(Z)}\Big] \quad (\text{琴生不等式}) \\
&= \underbrace{E_{q(Z)}\Big[lnP(X, Z)\Big] - E_{q(Z)}\Big[lnq(Z)\Big]}_{ELOB}
\end{aligned}
$$

我们想让分布 $q(Z)$ 与后验分布 $P(Z|X)$ 尽可能相似，当两个分布相同时，KL 散度为 0，ELBO 达到最大与似然函数 $lnP(X)$ 相等。所以目标就找到一个分布 $q(Z)$ 使得 ELBO 最大化。

选择一个简单的分布 q(Z)，每一个维度都统计独立:

$$q(Z) = \prod_{i=1}^{M} q_i(Z_i)$$

代入 $q(Z)$,ELOB 就改写为:

$$L(q) = \underbrace{\int_Z \prod_{i=1}^{M} q_i(Z_i) lnP(X,Z) dZ}_{part_1} - \underbrace{\int_Z \prod_{i=1}^{M} q_i(Z_i) \sum_{i=1}^{M} lnq_i(Z_i) dZ}_{part_2}$$

先化简 part-1:

$$\begin{aligned}
part_1 &= \int_{Z_1}\int_{Z_2}\cdots\int_{Z_M} \prod_{i=1}^{M} q_i(Z_i) lnP(X,Z) dZ_1 dZ_2 \cdots dZ_M \\
&= \int_{Z_j} q_j(Z_j) \left(\int_{i\neq j} \cdots \int \prod_{i\neq j}^{M} q_i(Z_i) lnP(X,Z) \prod_{i\neq j}^{M} dZ_i \right) dZ_j \\
&= \int_{Z_j} q_j(Z_j) \left(\int_{i\neq j} \cdots \int lnP(X,Z) \prod_{i\neq j}^{M} q_i(Z_i) dZ_i \right) dZ_j \\
&= \int_{Z_j} q_j(Z_j) \left[E_{\prod_{i\neq j}^{M} q(Z_i)} \left[lnP(X,Z) \right] \right] dZ_j
\end{aligned}$$

part-2 部分:

$$\begin{aligned}
part_2 &= \int_Z \prod_{i=1}^{M} q_i(Z_i) \sum_{i=1}^{M} lnq_i(Z_i) dZ \\
&= \int_Z \prod_{i=1}^{M} q_i(Z_i) [logq_1(Z_1) + logq_2(Z_2) + \cdots + logq_M(Z_M)] dZ \\
&= \underbrace{\int_Z \prod_{i=1}^{M} q_i(Z_i) logq_1(Z_1) dZ}_{\text{第一项为例}} + \cdots \\
&= \int_{Z_1 Z_2 \cdots Z_M} q_1(Z_1) q_2(Z_2) \cdots q_M(Z_M) logq_1(Z_1) dZ_1 dZ_2 \cdots dZ_M + \cdots \\
&= \int_{Z_1} q_1(Z_1) logq_1(Z_1) dZ_1 \cdot \underbrace{\int_{Z_2 \cdots Z_M} q_2(Z_2) \cdots q_M(Z_M) dZ_2 \cdots dZ_M}_{=1} + \cdots \\
&= \sum_{i=1}^{M} \left(\int_{Z_i} q_i(Z_i) lnq(Z_i) dZ_i \right)
\end{aligned}$$

推导参考 EM 的 Q 函数化简。如果我们只对 Z_j 感兴趣:

$$part_2 = \int_{Z_j} q_j(Z_j) lnq(Z_j) dZ_j + C$$

所以将化简后的两部分合起来:

$$L(q_j) = \int_{Z_j} q_j(Z_j) \left[\underbrace{E_{\prod_{i\neq j}^{M} q_i(Z_i)} \left[lnP(X,Z) \right]}_{ln\tilde{p}(X,Z_j)} \right] dZ_j - \int_{Z_j} q_j(Z_j) lnq_j(Z_j) dZ_j + C$$

$$
\begin{aligned}
L(q_j) &= \int_{Z_j} q_j(Z_j) ln \frac{ln\tilde{p}(X, Z_j)}{lnq_j(Z_j)} dZ_j + C \\
&= -KL(q_j \| \tilde{p}(X, Z_j)) \leq 0
\end{aligned}
$$

6.3 应用

接下来对下标做一下规范，用上标来表示第几个样本，下标表示变量的维度。

x 观测样本 $\rightarrow X = \{x^{(i)}\}_{i=1}^{N}$

z 隐变量 $\rightarrow Z = \{z^{(i)}\}_{i=1}^{N}$

所以似然函数：

$$
logP_\theta(X) = \sum_{i=1}^{N} logP_\theta(x^{(i)})
$$

单个样本的似然函数又可写成：

$$
logP_\theta(x^{(i)}) = \underbrace{ELBO}_{L(q)} + \underbrace{KL(q||p)}_{\geq 0}
$$

最大化数据集的似然函数，就要最大化每个样本的似然函数，目标函数为：

$$
\hat{q} = arg \min_q KL(q||p) = arg \max_q L(q)
$$

通过平均场理论，得到假设：

$$
q(z) = \prod_{i=1}^{M} q_i(z_i)
$$

其中的 z_i 可认为是最大团的概念。

迭代式:

$$
\begin{aligned}
logq_j(z_j) &= E_{\prod_{i \neq j} q_i(z_i)} \Big[logP_\theta(x^{(i)}, Z) \Big] + C \\
&= \int_{q_{i \neq j}} q_1 q_2 \cdots q_{j-1} q_{j+1} \cdots q_M \Big[logP_\theta(x^{(i)}, Z) \Big] dq_{i \neq j}
\end{aligned}
$$

比如在求 q_1 时，固定了除 q_1 以外其他参数不动，在求下一个 q_2 时，，用新的 q_1 替换掉旧的，且固定其他参数不动。不断迭代（坐标上升）。但是基于平均场的方法，对很多问题还是无法求解的，因为它的假设太强了。

6.4 SGVI

求后验分布 $q(Z|X)$ 等价求其分布的参数 ϕ, 所以更新公式可以改写为：

$$
\hat{\phi} = arg \max_\phi \underbrace{L(\phi)}_{ELBO}
$$

$$
\begin{aligned}
ELBO &= E_{q_\phi(Z)} \Big[\frac{log\tilde{p}(x^{(i)}, Z)}{logq_\phi(Z)} \Big] \\
&= E_{q_\phi(Z)} \Big[log\tilde{p}(x^{(i)}, Z) - logq_\phi(Z) \Big] \\
&= L(\phi)
\end{aligned}
$$

接下来求 $L(\phi)$ 关于 ϕ 的梯度：

$$\begin{aligned}\nabla_\phi L(\phi) &= \nabla_\phi E_{q_\phi(Z)}\Big[log\tilde{p}(x^{(i)},Z) - logq_\phi(Z)\Big] \\ &= \nabla_\phi \int_{q_\phi(Z)} q_\phi(Z)\Big[log\tilde{p}(x^{(i)},Z) - logq_\phi(Z)\Big]dZ \\ &= \underbrace{\int \nabla_\phi q_\phi(Z)\Big[log\tilde{p}(x^{(i)},Z) - logq_\phi(Z)\Big]dZ}_{part_1} + \underbrace{\int q_\phi(Z)\nabla_\phi\Big[log\tilde{p}(x^{(i)},Z) - logq_\phi(Z)\Big]dZ}_{part_2}\end{aligned}$$

先看第二项：

$$\begin{aligned}part_2 &= \int q_\phi(Z)\nabla_\phi\Big[log\tilde{p}(x^{(i)},Z) - logq_\phi(Z)\Big]dZ \\ &= -\int q_\phi(Z)\nabla_\phi logq_\phi(Z)dZ \\ &= -\int q_\phi(Z)\frac{1}{q_\phi(Z)}\nabla_\phi q_\phi(Z)dZ \\ &= -\nabla_\phi \int q_\phi(Z)dZ = -\nabla_\phi 1 = 0\end{aligned}$$

所以梯度就剩下第一项：

$$\begin{aligned}\nabla_\phi L(\phi) &= \int_{q_\phi(Z)} \nabla_\phi q_\phi(Z)\Big[log\tilde{p}(x^{(i)},Z) - logq_\phi(Z)\Big]dZ \\ &= \int_{q_\phi(Z)} q_\phi(Z)\nabla_\phi logq_\phi(Z)\Big[log\tilde{p}(x^{(i)},Z) - logq_\phi(Z)\Big]dZ \\ &= E_{q_\phi(Z)}\left[\nabla_\phi \underbrace{logq_\phi(Z)}_{\text{不稳定}}\Big[log\tilde{p}(x^{(i)},Z) - logq_\phi(Z)\Big]\right])\end{aligned}$$

但是如果 MCMC 采样到的概率 $q_\phi(Z)$ 特别小，在 log 函数中，这个值就会非常大，造成我们要求期望的量（梯度）的方差特别大，(本来求梯度就是为了逼近我们要求的分布，现在连梯度都不稳定了) 这样就需要更多的样本才能更好地近似。所以就不能直接用 MCMC 采样。

重参数化技巧：

6.5 指数族分布 + 变分推断

变分推断中的 ELBO（假设有两部分未知参数）：

$$L(q(Z,\beta)) = E_{q(Z,\beta)}[logP(X,Z,\beta)] - E_{q(Z,\beta)}[logq(Z,\beta)]$$

分别写出两个参数的指数族分布形式的似然函数：

$$P(\beta|Z,X) = h(\beta)exp\{\eta(Z,X)^T\varphi(\beta) - A_g(\eta(Z,X))\}$$

$$P(Z|\beta,X) = h(Z)exp\{\eta(\beta,X)^T\varphi(Z) - A_l(\eta(\beta,X))\}$$

这里将参数向量 η 看做是 Z,X 或 β,X 的函数。我们用两个分布去逼近上面两个似然：

$$q(\beta|\lambda) = h(\beta)exp\{\lambda^T\varphi(\beta) - A_g(\lambda)\}$$

$$q(Z|\phi) = h(Z)exp\{\phi^T\varphi(Z) - A_g(\phi)\}$$

我们不断调整参数 λ, ϕ，使得这两个分布与似然函数足够接近。所以，我们可以将 ELBO 看做是 λ, ϕ 的函数。

$$L(\lambda, \phi) = E_{q(Z,\beta)}[logP(X, Z, \beta)] - E_{q(Z,\beta)}[logq(Z, \beta)]$$

坐标上升法：

假设：$q(Z, \beta) = q(Z)q(\beta)$

1. 固定 ϕ, 优化 λ:

$$\begin{aligned} L(\lambda, \phi) &= E_{q(Z,\beta)}[logP(\beta|Z, X) + \underbrace{logP(Z|X)}_{\text{与}\lambda\text{无关}}] - E_{q(Z,\beta)}[logq(\beta)] - \underbrace{E_{q(Z,\beta)}[logq(Z)]}_{\text{与}\lambda\text{无关}} \\ &= E_{q(Z,\beta)}[logP(\beta|Z, X)] - E_{q(Z,\beta)}[logq(\beta|\lambda)] \end{aligned}$$

代入似然函数与我们给出的近似的分布:

$$\begin{aligned} &= \underline{E_{q(Z,\beta)}[logh(\beta)]} + E_{q(Z,\beta)}[\eta(Z, X)^T\varphi(\beta)] - \underline{E_{q(Z,\beta)}[A_g(\eta(Z, X))]} \\ &\quad - \underline{E_{q(Z,\beta)}[logh(\beta)]} - E_{q(Z,\beta)}[\lambda^T\varphi(\beta)] + E_{q(Z,\beta)}[A_g(\lambda)] \\ &= E_{q(Z)}[\eta(Z, X)^T] \cdot \underbrace{E_{q(\beta)}[\varphi(\beta)]}_{A_g^{\cdot}(\lambda)} - \underbrace{E_{q(\beta)}[\lambda^T\varphi(\beta)]}_{\lambda^T A_g^{\cdot}(\lambda)} + A_g(\lambda) \end{aligned}$$

上式对 λ 求导:

$$\begin{aligned} \frac{\partial L(\lambda, \phi)}{\partial \lambda} &= E_{q(Z)}[\eta(Z, X)^T] \cdot A_g^{\cdot\cdot}(\lambda) - A_g^{\cdot}(\lambda) - \lambda^T A_g^{\cdot\cdot}(\lambda) + A_g^{\cdot}(\lambda) \\ &= \left[E_{q(Z)}[\eta(Z, X)] - \lambda\right]^T A_g^{\cdot\cdot}(\lambda) = 0 \end{aligned}$$

我们得到下式，λ 的一步优化就完成了。

$$\lambda = E_{q(Z|\phi)}[\eta(Z, X)]$$

同理，ϕ 的优化:

$$\phi = E_{q(\beta|\lambda)}[\eta(X, \beta)]$$

7 马尔科夫链蒙特卡洛方法

共轭分布

分布复杂，找不到分布的特征，如均值。(分布的表达式是有的)

积分太难算:

$$E_{p(\theta|X)}[\theta] = \int_\theta \theta p(\theta|X)d\theta$$

从后验分布中采集样本，$\theta^{(i)}$ $P(\theta|X)$，去近似期望：

$$\hat{E}(\theta) = \frac{1}{N}\sum \theta^{(i)}$$

问题是如何从复杂的分布中取采样。从分布函数采样与从概率密度函数采样是等价的。但是并不是每个概率函数都可以求得分布函数。

https://blog.csdn.net/u011332699/article/details/74298555

7.1 cdf 采样

Inverse Sampling

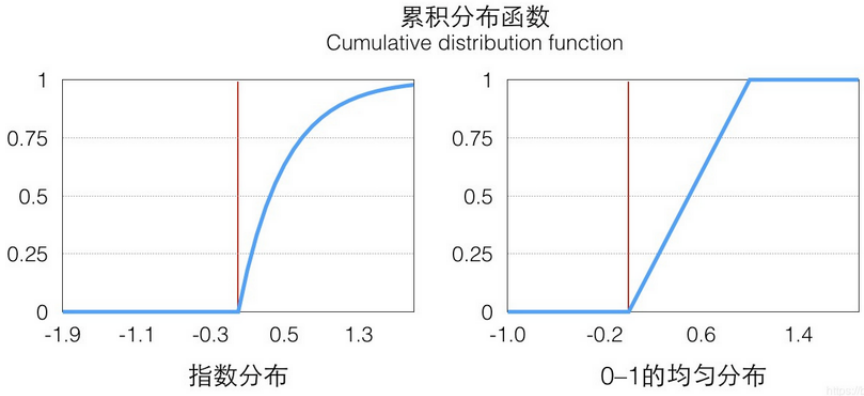
对于一些特殊的概率分布函数，比如指数分布：

$$p_{exp}(x) = \begin{cases} \lambda e^{-\lambda x}, x \geqslant 0 \\ 0, x < 0 \end{cases}$$

我们可以定义它的概率累积函数(Cumulative distribution function)，也就是(ps.这个’F’和前面的’f’函数并没有关系)

$$F(x) = \int_{-\infty}^{x} p(x)dx$$

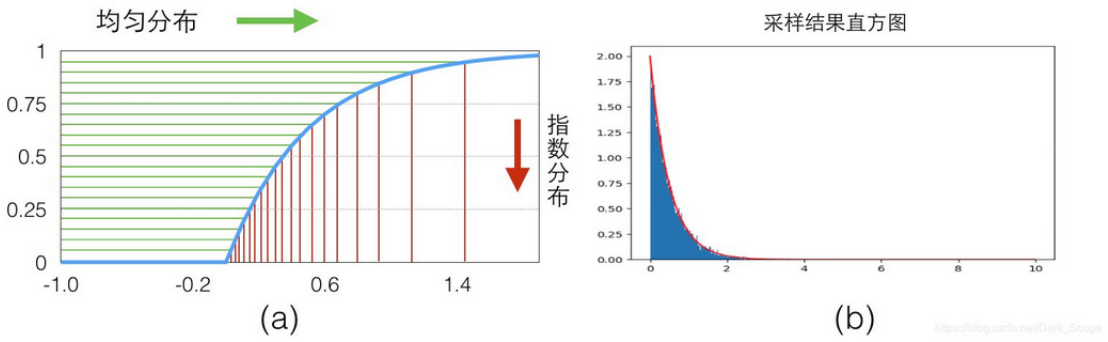
从图像上看就是概率密度函数小于x部分的面积。这个函数在x≥0的部分是一个单调递增的函数(在定义域上单调非减)，定义域和值域是[0,+∞)→[0,1)，画出来大概是这样子的一个函数，在p(x)大的地方它增长快（梯度大），反之亦然：



根据F(x)的定义，**它是exp分布的概率累积函数**，所以上面这个公式的意思是F－1(a)符合exp分布,**我们通过F的反函数将一个0到1均匀分布的随机数转换成了符合exp分布的随机数**，注意，以上推导对于cdf可逆的分布都是一样的，对于exp来说，它的反函数的形式是：

$$F_{exp}^{-1}(a) = -\frac{1}{\lambda}log(1-a)$$

具体的映射关系可以看下图(a)，我们从y轴0-1的均匀分布样本（绿色）映射得到了服从指数分布的样本（红色）。

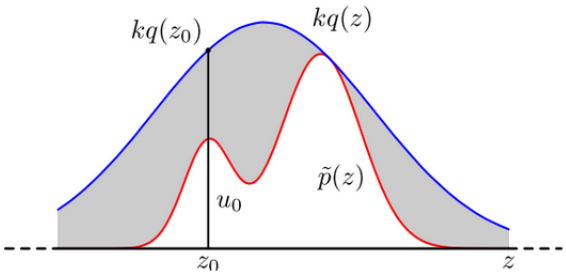


7.2 拒绝采样

1. 接受-拒绝采样

就算我们已知$p(x)$的分布，也很难得到一堆符合$p(x)$分布的样本$\{x_1, x_2, \ldots, x_n\}$来带入$g(x)$。

既然$p(x)$太复杂在程序中没法直接采样，那么我们设定一个程序可抽样的分布$q(x)$比如高斯分布，然后按照一定的方法拒绝某些样本，达到接近$p(x)$分布的目的，这就是**接受拒绝采样**。



接受拒绝采样具体操作如下，设定一个方便抽样的函数$q(x)$，以及一个常量k，使得已知的分布$p(x)$（红线）总在$kq(x)$（蓝线）的下方，

从方便抽样的$q(x)$分布抽样得到z_0

从均匀分布$(0, kq(z_0))$抽样得到u_0

如果u_0刚好落到灰色区域，拒绝这次采样，否则接受这次采样$x_t = z_0$

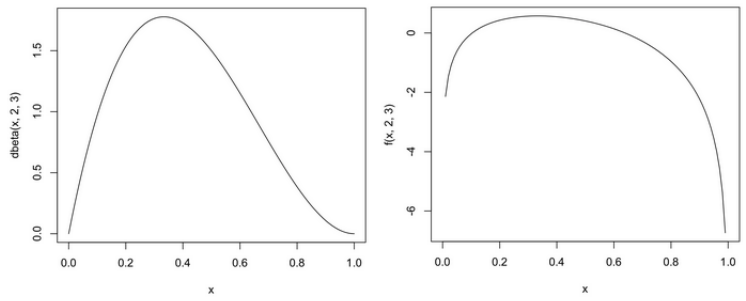
重复以上过程，得到接近$p(x)$分布的样本$\{x_1, x_2, \ldots, x_n\}$

在高维的情况下，接受-拒绝采样会出现两个问题，第一是合适的$q(x)$分布比较难以找到，第二是很难确定一个合理的 k 值。这两个问题会导致拒绝率很高，无用计算增加。

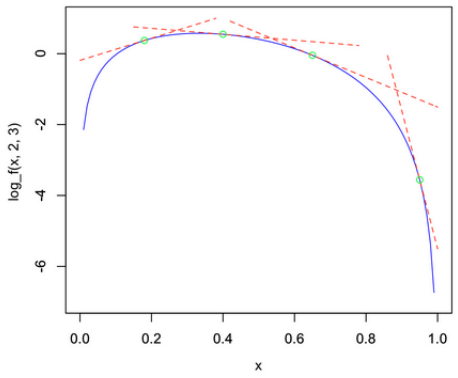
自适应的拒绝采样（Adaptive Rejection Sampling）

前面我们已经分析了，拒绝采样的弱点在于当被拒绝的点很多时，采样的效率会非常不理想。同时我们也支持，如果能够找到一个跟目标分布函数非常接近的参考函数，那么就可以保证被接受的点占大多数（被拒绝的点很少）。这样一来便克服了拒绝采样效率不高的弱点。如果函数是 log-concave 的话，那么我们就可以采样自适应的拒绝采样方法。什么是 log-concave 呢？还是回到我们之前介绍过的 Beta 分布的 PDF，我们用下面的代码来绘制 Beta(2, 3) 的函数图像，以及将 Beta(2, 3) 的函数取对数之后的图形。

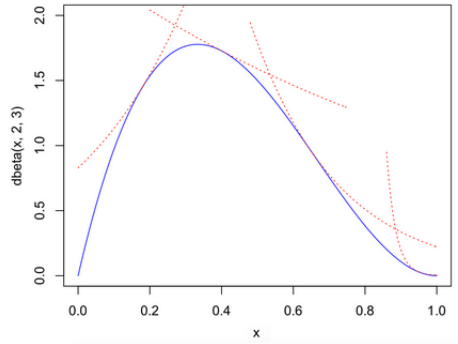
上述代码的执行结果如下所示。其中左图是Beta(2, 3) 的函数图像，右图是将 Beta(2, 3) 的函数取对数之后的图形，你可以发现结果是一个凹函数（concave）。那么Beta(2, 3) 就满足log-concave的要求。



然后我们在对数图像上找一些点做图像的切线，如下图所示。因为对数图像是凹函数，所以每个切线都相当于一个超平面，而且对数图像只会位于超平面的一侧。



再把这些切线转换回原始的Beta(2, 3)图像中，显然原来的线性函数会变成指数函数，它们将对应用下图中的一些曲线，这些曲线会被原函数的图形紧紧包裹住。特别是当这些的指数函数变得很多很稠密时，以彼此的交点作为分界线，我们其实相当于得到了一个分段函数。这个分段函数是原函数的一个逼近。用这个分段函数来作为参考函数再执行Reject Sampling，自然就完美的解决了我们之前的问题。



7.3 重要性采样

Importance Sampling

上面描述了两种从另一个分布获取指定分布的采样样本的算法，对于1.在实际工作中，一般来说我们需要sample的分布都及其复杂，不太可能求解出它的反函数，但p(x)的值也许还是可以计算的。对于2.找到一个合适的cq(x)往往很困难，接受概率有可能会很低。

那我们回过头来看我们sample的目的：其实是想求得E[f(x)],x～p,也就是

$$E[f(x)] = \int_x f(x)p(x)dx$$

如果符合p(x)分布的样本不太好生成，我们可以引入另一个分布q(x)，可以很方便地生成样本。使得

$$\int_x f(x)q(x)dx = \int_x f(x)\frac{p(x)}{q(x)}q(x)dx = \int_x g(x)q(x)dx,$$

$$where\, g(x) = f(x)\frac{p(x)}{q(x)} = f(x)w(x)$$

我们将问题转化为了求g(x)在q(x)分布下的期望！！！

我们称其中的 $w(x) = \frac{p(x)}{q(x)}$ 叫做**Importance Weight**.

在高维情况下，上面的方法都不太行。

7.4 马尔科夫链的平稳分布

在马尔科夫链中，如果其状态空间为 $\{1, 2, \cdots, k\}$，条件转移矩阵（随机矩阵，特征值的绝对值小于等于 1）：

$$\begin{pmatrix} Q_{11} & Q_{12} & \cdots & Q_{1k} \\ Q_{21} & Q_{22} & \cdots & Q_{2k} \\ \vdots & \vdots & \vdots & \vdots \\ Q_{k1} & Q_{k2} & \cdots & Q_{kk} \end{pmatrix}$$

则第 t+1 个时刻，样本 x=j 的概率：

$$q^{(t+1)}(x = j) = \sum_{i=1}^{k} q^{(t)}(x = i)Q_{ij}$$

把第 t+1 个时刻样本所有取值的概率写成一个向量：

$$\boldsymbol{q^{(t+1)}} = \begin{pmatrix} q^{(t+1)}(x = 1) & q^{(t+1)}(x = 2) \cdots q^{(t+1)}(x = k) \end{pmatrix}$$

代入可得：

$$\begin{aligned} \boldsymbol{q^{(t+1)}} &= \begin{pmatrix} \underbrace{\sum_{i=1}^{k} q^{(t)}(x = i)Q_{i1}}_{\text{每一列和}} & \sum_{i=1}^{k} q^{(t)}(x = i)Q_{i2} \cdots \sum_{i=1}^{k} q^{(t)}(x = i)Q_{ik} \end{pmatrix} \\ &= \boldsymbol{q^{(t)}Q} \end{aligned}$$

所以：

$$\boldsymbol{q}^{(t+1)} = \boldsymbol{q}^{(t)}\boldsymbol{Q} = \cdots = \boldsymbol{q}^{(1)}\boldsymbol{Q}^t$$

将随机矩阵特征值分解, 不妨设 $\lambda_i = 1$：

$$\boldsymbol{Q} = \boldsymbol{A}\boldsymbol{\Lambda}\boldsymbol{A}^{-1}$$

所以, 经过足够多的 m 次采样（收敛的时间叫做 mixing time)：

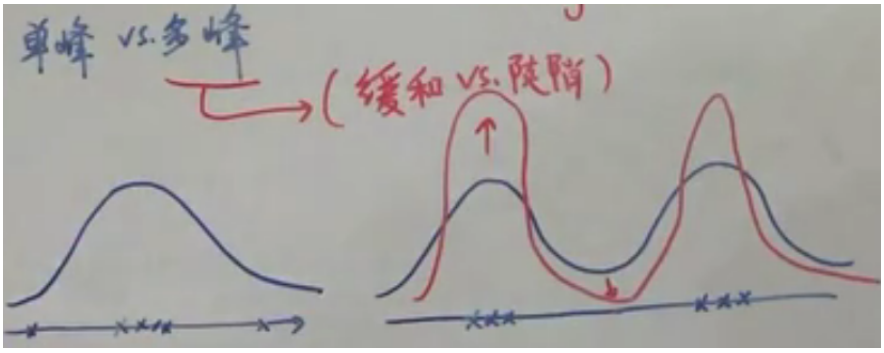
$$\boldsymbol{q}^{(t+1)} = \boldsymbol{q}^{(1)}\boldsymbol{A}\boldsymbol{\Lambda}^m\boldsymbol{A}^{-1}$$

$$s.t. \quad \boldsymbol{\Lambda}^m = \begin{pmatrix} 0 & & & & \\ & \ddots & & & \\ & & 1 & & \\ & & & \ddots & \\ & & & & 0 \end{pmatrix}$$

当 t>m 时, 后面的分布都相同了:

$$\boldsymbol{q}^{(m+1)} = \boldsymbol{q}^{(m+2)} = \cdots = \boldsymbol{q}^{(\infty)}$$

存在问题：理论上只能保证收敛性无法知道何时收敛；混合时间过长（高维，维度之间相关性强)；样本之间有一定的相关性。当分布是多峰的时候，样本很容易就只在一个峰里面打转，很难突破最低值到达另一个峰, 造成混合时间太长。



7.5 MCMC

马尔科夫链的时间和状态都是离散的，P 是状态转移矩阵。

$$\pi(x^*) = \int \pi(x)P(x \to x^*)dx$$

如果分布 $\pi(k)$ 满足上式，则称 $\pi(k)$ 为 X_t 的平稳分布。

满足细致平衡条件，则可推出 π 为平稳分布（两个点之间能跳过来也能跳过去)：

$$\pi(x)P(x \to x^*) = \pi(x^*)P(x^* \to x)$$

Metropolis Hasting：

1. initialise $x^{(0)}$
2. **for** i $= 0$ to $N-1$
   $u \sim U(0,1)$
   $x^* \sim q(x^*|x^{(i)})$
   if $u < \min\left(1, \frac{\pi(x^*)q(x|x^*)}{\pi(x)q(x^*|x)}\right)$
   $x^{(i+1)} = x^*$
   else
   $x^{(i+1)} = x^{(i)}$

- The take-home message here, is that it does not "disgard" samples like rejection sampling. It simply "repeats" samples.
- If the same sample repeats too many times, it has **bad mixing**

- $K(x \to x^*)$ includes the joint density of the following:
  1. Propose x^* from $q(x^*|x)$,
  2. then accept x^* with ratio $\alpha(x^*, x) = \min\left(1, \frac{\pi(x^*)q(x|x^*)}{\pi(x)q(x^*|x)}\right)$
- very easily verify it satisfy **detailed balance**:

$$\begin{aligned}
\pi(x)q(x^*|x)\alpha(x^*,x) &= \pi(x)q(x^*|x)\min\left(1, \frac{\pi(x^*)q(x|x^*)}{\pi(x)q(x^*|x)}\right) \\
&= \min\left(\pi(x)q(x^*|x), \pi(x^*)q(x|x^*)\right) \\
&= \pi(x^*)q(x|x^*)\min\left(1, \frac{\pi(x)q(x^*|x)}{\pi(x^*)q(x|x^*)}\right) \\
&= \pi(x^*)q(x|x^*)\alpha(x,x^*)
\end{aligned}$$

吉布斯采样：

Gibbs sampling algorithm:

- given a starting sample $(x_1, y_1, z_1)^\top$
- you want to sample

$$\{(x_2, y_2, z_2)^\top, (x_3, y_3, z_3)^\top, \ldots, (x_N, y_N, z_N)^\top\} \sim P(x, y, z)$$

- Then the algorithm goes:

$$\begin{aligned} x_2 &\sim P(x|y_1, z_1) \\ y_2 &\sim P(y|x_2, z_1) \\ z_2 &\sim P(z|x_2, y_2) \end{aligned}$$

$$\begin{aligned} x_3 &\sim P(x|y_2, z_2) \\ y_3 &\sim P(y|x_3, z_2) \\ z_3 &\sim P(z|x_3, y_3) \end{aligned}$$

吉布斯采样其实是 MH 采样的接受率为 1 的一种特殊情况。

Looking at the M-H accpetance ratio

- Let $\mathbf{x} = x_1, \ldots, x_D$.
- When sampling k^{th} component, $q_k(\mathbf{x}^*|\mathbf{x}) = \pi(x_k^*|\mathbf{x}_{-k})$
- When sampling k^{th} component, $\mathbf{x}^*_{-k} = \mathbf{x}_{-k}$

$$\frac{\pi(\mathbf{x}^*)q(\mathbf{x}|\mathbf{x}^*)}{\pi(\mathbf{x})q(\mathbf{x}^*|\mathbf{x})} = \frac{\pi(\mathbf{x}^*)\pi(x_k|\mathbf{x}^*{}_{-k})}{\pi(\mathbf{x})\pi(x_k^*|\mathbf{x}_{-k})} = \frac{\pi(x_k^*|\mathbf{x}^*_{-k})\pi(x_k|\mathbf{x}^*{}_{-k})}{\pi(x_k|\mathbf{x}_{-k})\pi(x_k^*|\mathbf{x}_{-k})} = 1$$

下标 X_{-k} 表示 X 的除去 k 的所以维度。
吉布斯采样就是把目标分布 P 对应的条件概率当做状态转移分布 Q。

采样的动机：

1. 采样本身就是常见的额任务

12. 求和或求积分

什么是好的样本？

11. 样本趋向于高概率区域

1 2. 样本之间相互独立

采样困难：（高维）

1. 归一化系数无法求解

2. 维度太高，无法直接采样（需知道每个状态的概率，要遍历）

拒绝采样与重要性采样都构造了一个分布 q，1. 要求 q 与 p 接近，2. 且 q 要更简单。